Languages in the European information society - German

Edited by: Burchardt, Aljoscha ; Egg, Markus ; Eichler, Kathrin ; Krenn, Brigitte ; Lessmöllmann, Annette ; Rehm, Georg ; Stede, Manfred ; Uszkoreit, Hans ; Volk, Martin

Abstract: Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.Innovative language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the Jeopardy game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:- Should our communications and knowledge infrastructure be dependent upon monopolistic companies?- Can we truly rely on language-related services that can be immediately switched off by others?- Are we actively competing in the global market for research and development in language technology?- Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?- Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?This whitepaper for the German language demonstrates that a lively language technology industry and research environment exists in Germany, Austria and Switzerland. Although a number of technologies and resources for Standard German exist, there are fewer technologies and resources for the German language than for the English language. The existing technologies and resources also have a poorer quality.According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the German language can be achieved.

META-NET White Paper Series

Languages in the European Information Society

– German –

Early Release Edition

META-FORUM 2011

27-28 June 2011

Budapest, Hungary


The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, journalists, politicians, language communities and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a European Commission Network of Excellence, has conducted an analysis of current language resources and technologies. This analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed, expert analysis and assessment of the current situation will help maximise the impact of additional research and minimize any risks.

META-NET consists of 47 research centres from 31 countries that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

META-NET
DFKI Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin
Germany

office@meta-net.eu
http://www.meta-net.eu

Authors

Dr. Aljoscha Burchardt, DFKI
Prof. Dr. Markus Egg, Humboldt-Universität zu Berlin
Kathrin Eichler, DFKI
Dr. Brigitte Krenn, ÖFAI
Prof. Dr. Annette Leßmöllmann, Hochschule Darmstadt
Dr. Georg Rehm, DFKI
Prof. Dr. Manfred Stede, Universität Potsdam
Prof. Dr. Hans Uszkoreit, Universität des Saarlandes and DFKI
Prof. Dr. Martin Volk, Universität Zürich

Table of Contents

Executive Summary

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.

Innovative language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.

Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the *Jeopardy* game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:

- Should our communications and knowledge infrastructure be dependent upon monopolistic companies?
- Can we truly rely on language-related services that can be immediately switched off by others?
- Are we actively competing in the global market for research and development in language technology?
- Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?
- Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?

This whitepaper for the German language demonstrates that a lively language technology industry and research environment exists in Germany, Austria and Switzerland. Although a number of technologies and resources for Standard German exist, there are fewer technologies and resources for the German language than for the English language. The existing technologies and resources also have a poorer quality.

According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the German language can be achieved.

A Risk for Our Languages and a Challenge for Language Technology

We are currently witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digitised and network communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts like Luther's translation of the Bible into common language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled an exchange across languages;
- the creation of journalistic and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology helped to automate and facilitate many of the processes:

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents faster than a fax machine;
- Skype makes Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick and approximate translations;
- social media platforms facilitate collaboration and information sharing.

Although such tools and applications are helpful, they currently cannot sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely.

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European

foreign ministers speak in their native language. We might want a platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

It is clear that communication needs have a different quality as compared to a few years ago. In a global economy and information space, more languages, speakers and content confront us and require us to quickly interact with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter and YouTube) is only the tip of the iceberg.

A global economy and information space confronts us with more languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.[1] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, "Which European languages will thrive and persist in the networked information and knowledge society?"

Which European languages will thrive and persist in the networked information and knowledge society?

Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also led to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe's multitude of languages is also a vital part of its social success.[2] While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost that would weaken Europe's global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[3]

The wide variety of languages in Europe is one of its most important cultural assets and an essential part of Europe's success.

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.[4] Yet, this existing capacity is not enough to satisfy current and future needs.

Language technology is a key enabling technology that can protect and foster European languages. Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates regardless of language barriers or computer skills. Language technology already assists everyday tasks, such as writing e-mails, conducting an online search or booking a flight. We benefit from language technology when we:

Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.

- find information with an Internet search engine;
- check spelling and grammar in a word processor;
- view product recommendations at an online shop;
- hear the verbal instructions of a navigation system;
- translate web pages with an online service.

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies for each language.

In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

Opportunities for Language Technology

Language technology can make automatic translation, content production, information processing and knowledge management possible for all European languages. Language technology can also further the development of intuitive language-based interfaces for household electronics, machinery, vehicles, computers and robots. Although many prototypes already exist, commercial and industrial applications are still in the early stages of development. Recent achievements in research and development have created a genuine window of opportunity. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

Language applications, voice-based user interfaces and dialogue systems are traditionally found in highly specialised domains, and they often exhibit limited performance. One active field of research is the use of language technology for rescue operations in disaster areas. In such high-risk environments, translation accuracy can be a matter of life or death. The same reasoning applies to the use of language technology in the health care industry. Intelligent robots with cross-lingual language capabilities have the potential to save lives.

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union that makes both economic and cultural sense. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for European can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. We cannot wait ten or twenty years for significant improvements to be made that can further communication and productivity in our multilingual environment.

The current pace of technological progress is too slow to arrive at substantial software products within the next ten to twenty years.

Language technologies with broad use, such as the spelling and grammar features in word processors, are typically monolingual, and they are only available for a handful of languages. Applications for multilingual communication require a certain level of sophistication. Machine translation and online services like Google Translate or Bing Translator are excellent at creating a good approximation of a document's contents. But such online services and professional MT applications are fraught with various difficulties when highly accurate and complete translations are required. There are many well-known examples of funny sounding mistranslations, for example, literal translations of the names *Bush* or *Kohl*, that illustrate the challenges language technology must still face.

Language Acquisition

To illustrate how computers handle language and why language acquisition is a very difficult task, we take a brief look at the way humans acquire first and second languages, and then we sketch how machine translation systems work—there's a reason why the field of language technology is closely linked to the field of artificial intelligence.

Humans acquire language skills in two different ways. First, a baby learns a language by listening to the interaction between speakers of the language. Exposure to concrete, linguistic examples by language users, such as parents, siblings and other family members, helps babies from the age of about two or so produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning languages.

Humans acquire language skills in two different ways: learning examples and learning the underlying language rules.

Learning a second language usually requires much more effort when a child is not immersed in a language community of native speakers. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

The two main types of language technology systems acquire language capabilities in a similar manner as humans. Statistical approaches obtain linguistic knowledge from vast collections of concrete example texts in a single language or in so-called parallel texts that are available in two or more languages. Machine learning algorithms model some kind of language faculty that can derive patterns of how words, short phrases and complete sentences are correctly used in a single language or translated from one language to another. The sheer number of sentences that statistical approaches require is huge. Performance quality increases as the number of analyzed texts increases. It is not uncommon to train such systems on texts that comprise millions of sentences. This is one of the reasons why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, available online information, and translation services such as Google Search and Google Translate rely on a statistical (data-driven) approach.

The two main types of language technology systems acquire language in a similar manner as humans.

Rule-based systems are the second major type of language technology. Experts from linguistics, computational linguistics and computer science encode grammatical analysis (translation rules) and compile vocabulary lists (lexicons). The establishment of a rule-based system is very time consuming and labour intensive. Rule-based systems also require highly specialised experts. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The advantage of rule-based systems is that the experts can more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. Due to financial constraints, rule-based language technology is only feasible for major languages.

German in the European Information Society

General Facts

With about 100 million native speakers, German is the most widely spoken native language in the European Union. It is the commonly used language in Germany, Austria and Liechtenstein and one of the official languages in Switzerland, Luxembourg and Belgium, where it is used by parts of the population. Around the world, German is also spoken by around 30 million non-native speakers[5] and ranks second as foreign language studied in the EU, after English[6].

In Germany, it is the common spoken and written language and the native language of the vast majority of the population. Minority languages in the sense of the European Charter on Regional and Minority Languages are Danish and North Frisian in Schleswig-Holstein, Upper Sorbian in Saxony, Lower Sorbian in Brandenburg, Saterland Frisian in Lower Saxony and the Romani language of the German Roms and Sinti throughout the country, each of these groups representing some tens to hundreds of thousands of speakers.[7] In addition, there exist immigrant languages, mainly Turkish with roughly 3.3 million speakers in Germany.

In Austria and Liechtenstein, German is the official and most common spoken and written language. In Austria, recognized minority languages are Slovenian, Croatian (Burgenland-Kroatisch), Slovacian, Romani, Hungarian, and Czech. Other languages spoken in Austria are Turkish and languages of former Yugoslavia including Bosnian, Croatian, and Serbian.

In Switzerland, German shares its official status with French, Italian, and Rhaeto-Romanic, in Belgium with Dutch and French, and in Luxembourg with French and Luxembourgish. German variants are also spoken by minorities in other EU countries, e.g., France (Alsace, Lorraine), Italy (South Tyrol), and Poland (Silesia).

German has a large variety of dialects, e.g. Bavarian and Swabian. By and large, they all underlie the same grammar, even though some dialects exhibit slightly different syntactic constructions. The German separation is still reflected in some lexical differences such as *Plastik* vs. *Plaste* ('plastics').

Austrian German (AT) is a variant of German (Oberdeutsch) which mainly differs from the official language used in Germany (DE) in parts of its lexicon, e.g., chair: *Sessel* (AT) versus *Stuhl* (DE), arm chair: *Fauteuil* (AT) versus *Sessel* (DE), tobacconist: *Trafik* (AT) versus *Tabakladen* (DE). Also, lenisation is widespread in spoken Austrian German. For instance, there is no pronounced phonemic distinction between *backen* (to bake) and *packen* (to pack) or *Teich* (lake) and *Deich* (dyke). In addition, other than in Germany Standard German, past tense is almost unused in Austrian German. Instead, perfect tense is employed to express past events.

Swiss German borrowed some French words, e.g., bicycle: *Velo* (CH) versus *Fahrrad* (DE). There are also some morphological and orthographical variances.[8] For example, 'ss' is used instead of 'ß' and some words are spelled differently, e.g., cereal: *Müesli* (CH) instead of *Müsli* (DE). In Switzerland, with four official languages, multilingualism is a matter of course.

Particularities of the German Language

German exhibits some specific characteristics, which contribute to the richness of the language but are challenges for computational processing of natural language. Some of these characteristics allow the speakers to express ideas in a wide variety of ways. First, word order is relatively free in German sentences. Consider, e.g., the English sentence

The woman gave the man an apple.

In English, there are two more ways to express the same idea, namely:

The woman gave an apple to the man.

An apple was given to the man by the woman.

In German, there exist at least nine possible ways (even though some of them would hardly be used):

Die Frau gab dem Mann einen Apfel.

Einen Apfel gab die Frau dem Mann.

Dem Mann gab die Frau einen Apfel.

Ein Apfel wurde dem Mann von der Frau gegeben.

Dem Mann wurde von der Frau ein Apfel gegeben.

Dem Mann wurde ein Apfel von der Frau gegeben.

Ein Apfel wurde von der Frau dem Mann gegeben.

Von der Frau wurde dem Mann ein Apfel gegeben.

Von der Frau wurde ein Apfel dem Mann gegeben.

Second, German is extremely productive when it comes to coining new words. This is mainly due to the compounding system, which allows speakers to put together words (and affixes) in a quite simple way. In theory, this allows the creation of infinitely long words:

Verteidigung (defence)

Verteidigungsminister (minister of defence)

Selbstverteidigungsminister (minister of self-defence)

Bundesselbstverteidigungsminister (federal minister of self-defence)

Usually, humans can easily derive the meaning of these neologisms, but machines have difficulties processing them.

Other specific characteristics of German that make automatic processing of German difficult are the tendency to use comparably long and nested sentences, and separable verb prefixes that can occur far away from the verb in sentences like

*Er **stellte** sich, nachdem er mir ein Getränk angeboten hatte und wir ins Gespräch gekommen waren, **vor**.*

*(He **introduced** himself after he had offered me a drink and we had started a conversation.)*

The difference in meaning between verbs that are built with different prefixes like *vor*, *ein*, or *aus* can be confusing for learners of German. E.g., the verb *stellen* (to put) appears in *vorstellen* (imagine, introduce), *einstellen* (hire, discontinue, regulate), *ausstellen* (exhibit, switch off, issue), and many other verbs.

Recent developments

From the 1950s on, American television series and movies began to dominate the German market. Even though foreign films and series are usually dubbed into German (in contrast to many other countries such as Sweden and Poland), the strong presence of the American way of life in the media influenced the German culture and language. Due to the continuing triumph of English and American music since the 1960s, Germans have been exposed to a lot of English during their adolescence for generations. English soon acquired the status of a 'cool/hip' language, which it has kept up to the present day.

This continuing status is reflected by the sheer number of present-day loan words from English (so-called anglicisms). A systematic investigation of neologisms in German newspapers since 2000 revealed that about one third of these neologisms are complete or partial anglicisms.[9] In most cases these words fill some gap in the vocabulary, i.e., they complement native German words rather than competing with them.

However, in some areas, anglicisms have started to replace existing German vocabulary. One example is the use of English titles in job advertisements, in particular for executive positions, e.g. 'Human Resource Manager' instead of *Personalleiter*. A strong tendency to overuse anglicisms can also be detected in product advertisements. In 2003, a study on English slogans used for advertising by German companies was carried out by Endmark and revealed that almost all of the 12 investigated slogans were misunderstood by more than half of the respondents, which eventually induced the companies to replace their slogans by German equivalents. The example demonstrates the importance of raising awareness for a development that runs the risk of excluding large parts of the population from taking part in information society, namely those who are not familiar with English.

Language cultivation in Germany

Lacking linguistic legislation, Germany has no institutional body with responsibility for developing or implementing any given policy. However, there are a number of non-governmental but publicly funded organisations which play an active role in promoting the German language. The Goethe Institute works in partnership with the Federal Foreign Affairs Ministry and offers German language courses all over the world in order to strengthen the international position of the German language. Other organizations, which aim at raising language awareness and promoting German language culture in Germany, include the German Academy for Language and Literature (DASD) and the Society for the German Language (GfDS), which has been charged by the Federal Parliament with controlling legislative texts. The Institute for the German Language (IDS) is the central research centre for German.

In addition, individual authors contribute to linguistic awareness by discussing undesirable developments such as the influential use of incorrect apostrophes (e.g. *'Maria's Haus'*, correct: *Marias Haus*), business speech, or neologisms. The most well-known

author of this genre is probably Bastian Sick with his column Zwiebelfisch[10]. Private initiatives specifically turn against anglicisms: The initiative *Verein Deutsche Sprache* annually awards the *Kulturpreis Deutsche Sprache* for creative contributions to the development of the German language (e.g. this year to singer Udo Lindenberg), the campaign *Aktion lebendiges Deutsch* regularly organizes contests for Germanizing anglicisms.

However, Germany does not maintain a language academy prescribing the 'correct' usage of the language (like, e.g., the *Académie Francaise* in France or the *Academía Real* in Spain). The *Duden* dictionary used to be the prescriptive source for the spelling and grammar of German, but nowadays it pursues a more descriptive approach.[11]

Political efforts concerned with the German language are rare. In 2006, Austria, Germany, Liechtenstein and Switzerland agreed on an orthography reform after ten years of discussion. The original reform was modified and weakened, allowing writers more freedom. The new spelling conventions were not accepted universally; many large newspapers and publishers use a mixture of old and new spelling conventions ('house orthography').

There are almost no measures to protect the status of the German language. A recent initiative (Dec. 2008), started by several politicians and private associations, most notably the Verein Deutsche Sprache e.V., calls for a change to the constitution to allow a clause to be added defining German as 'the language of the Federal Republic of Germany'. This demand was turned down by the Federal Parliament, but is still the subject of lively debates and remains a hot topic. Also, a radio quota regulating the percentage of music sung in German, comparable to the one in France, was considered in 2004, but never passed.

The examples above illustrate the disadvantageous situation of the German language compared to, e.g., French, which is strongly promoted by the global community of French-speaking peoples in the so-called Francophonie. This comparably low level of cultural identity associated with the German language encourages an attitude of tolerance and openness towards cultural diversity, but can also pose a threat to maintaining high standards for German.

Language in Education

The first PISA study, conducted in 2000, revealed that German students performed below OECD average with respect to reading literacy. Students with migration background received particularly low results. The ensuing debate has increased public awareness for the importance of language learning, especially with respect to integration.

Following the recommendations of the OECD, Germany has adopted several laws on early language training in the last decade. One example is the *Gesetz zur vorschulischen Sprachförderung* (Law for promoting pre-school language learning), which came into effect in April 2008 in Berlin, a city with a very high rate of children whose native tongue is not German (> 90% in some areas in the district of Neukölln). The law introduces a compulsory German test for children not attending kindergarten before school enrolment and offers enhanced language training for those with insufficient German language skills.

Steps like this have proven successful: The PISA study in 2009 showed that the reading literacy of Germans has improved significantly since 2000, in particular, the one of children with migration background. However, the differences in language skills between students of German and non-German background are still large compared to those in other countries.[12] The differences are particularly obvious in Austria, which is among the three OECD countries which show the biggest gap between native Austrian youth and youth with migrant background with regard to reading literacy.

The German language has become a central part of the immigration policy in Germany. The law on controlling and limiting immigration and regulating the stay and integration of migrants, which came into force in 2005, places particular emphasis on learning German through integration classes. These classes include 600 hours of language teaching plus 30 hours of introduction to German history, culture and law and order. Those immigrants who do not take part in the integration classes risk having their state benefits reduced and may encounter difficulties when it comes to renewing their residence permit. Participation in the course is in any case a prerequisite for obtaining permission to stay permanently in Germany.

Language skills are the key qualification needed in education as well as for personal and professional communication. Still, the status of German as school subject in higher education is comparably low. According to OECD figures published in 2003[13], German language teaching makes up about 20% of the school lessons of 9-to-11-year-old students, compared to almost one third of native language lessons in France, Greece and the Netherlands.

Increasing the quantity of German language teaching in schools is one possible step towards providing students with the language skills required for an active participation in society. language technology can make an important contribution here by offering so-called computer-assisted language learning (CALL) systems, which allow students to experience language in a playful way, for example by linking special vocabulary in electronic text to comprehensible definitions or to audio or video files supplying additional information, e.g., the pronunciation of a word.

International aspects

Germany is often referred to as the land of *Dichter und Denker* – of poets and thinkers. German contributions to the fields of literature, philosophy and science have been immense. The works by authors like Goethe, Kafka, and Hesse have gained international fame; the philosophies of Kant, Hegel, Marx, Nietzsche and the theory of psychoanalysis by Freud have a lasting impact on modern society. Scientists from the German-speaking countries have won numerous Nobel prizes in literature, economy, physics, chemistry, and medicine.

At the beginning of the 20th century, the German-speaking countries were at the forefront of scientific disciplines and German was the major scientific language, with 30% of scientific publications written in German. Since then, the importance of German as a scientific language has decreased dramatically: Nowadays, less than 5% of scientific publications are written in German, most of them in disciplines like law, philosophy, and theology.[14] This can only partly be attributed to a decline in scientific contributions from German-speaking countries. Even in universities of these

countries, German is strongly rivalled or has been overtaken by English as publication language in many disciplines.

Similarly, this is true of the business world. In many large and internationally active companies, English has become the *lingua franca*, both in written (emails and documents) and oral communication (e.g. talks). These developments strongly affect the status of German as a foreign language. Pragmatic reasons for learning German (e.g. better chances on the job market) have lost importance. German is more and more losing out to English (and recently Chinese).

Within the European Union, German officially has a high status, being one of the three working languages of the European Commission, along with English and French. In practice however, German is hardly used in European Union business. Only 3% of the documents sent by the European Commission to member states are written in German.[15] First political actions to face this problem have been taken. In 2006, President of the *Bundestag*, Norbert Lammert, wrote a letter to EU commission saying that the Bundestag will reject to deal with contracts and other relevant documents if a German translation to the texts is missing.

Language technology can address this challenge from a different perspective by offering services like Machine Translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

German on the Internet

In 2010, almost 70% of the Germans and 74.8% of the Austrian population were Internet users[16],[17] Most of them said they were online every day. Among young people, the proportion of users in both countries is even higher. The existence of an active German-speaking web community is also mirrored by the fact that the German Wikipedia is the second largest Wikipedia after English (not counting automatically translated versions such as the Thai Wikipedia).

With about 14 million Internet domains in November 2010[18], Germany's top-level country domain .de is the world's largest country extension and second only to the extension .com.[19] This dominant Internet presence suggests that there is a vast amount of German language data available on the web. In addition, some bi-lingual resources like the online dictionary LEO[20] are freely available.

For language technology, the growing importance of the Internet is important in two ways. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas for language technology.

The most commonly used web application is certainly Web Search, which involves the automatic processing of language on multiple levels, as we will see in more detail the second part of this paper. It involves sophisticated language technology, differing for each language. For German, this comprises matching 'ä' and 'ae' or taking into account the use of upper case to distinguish, e.g., between the noun *Fliegen* (flies) and the verb *fliegen* (to fly).

It is an expressed political aim in Germany and other European countries to ensure equal opportunities for everyone. In particular,

the “Gesetz zur Gleichstellung behinderter Menschen” (Law on Equal Opportunities for the Disabled), which came into force in 2002, addresses the issue of “Barrierefreie Informationstechnik” (barrier-free information technology), stating that public agencies need to make sure that their web sites and internet services can be used by the disabled without restrictions. User-friendly language technology tools offer the principal solution to satisfy this regulation, for example by offering speech synthesis for the blind.

Internet users and providers of web content can also profit from language technology in less obvious ways, e.g., if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, comparatively little usable language technology is developed and applied, compared to the anticipated need. This may be due to the complexity of the German language and the number of technologies involved in typical language technology applications. In the next chapter, we will present an introduction to language technology and its core application areas as well as an evaluation of the current situation of language technology support for German.

Selected Further Reading

Lothar Lemnitzer: Von Aldianer bis Zauselquote, Gunter Narr Verlag, Tübingen, 2007.

Wolf Schneider: Speak German! – Warum Deutsch manchmal besser ist. Rowohlt, Reinbek bei Hamburg, 2008.

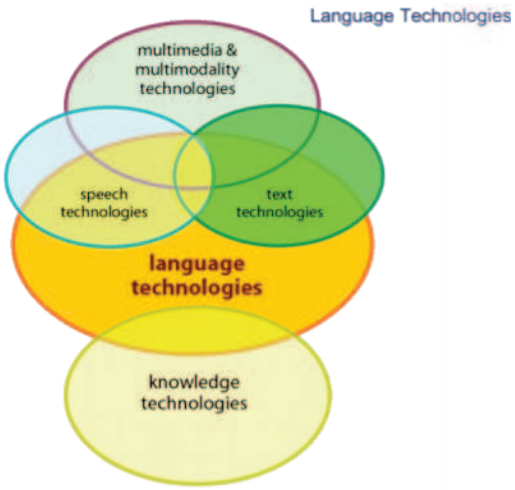
Bastian Sick: Der Dativ ist dem Genitiv sein Tod – Ein Wegweiser durch den Irrgarten der deutschen Sprache. Kiepenheuer und Witsch, Köln, 2004.

Thomas Steinfeld: Der Sprachverführer. Die deutsche Sprache: was sie ist, was sie kann. Carl Hanser Verlag, München, 2010.

Language Technology Support for German

Language Technologies

Language technologies are information technologies that are specialized for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under either speech or text technologies. Among those are technologies that link language to knowledge. The figure on the right illustrates the Language Technology landscape. In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language and spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents.



Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. The figure on the right displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the structure and meaning of the text input:

and meaning of the text input:

- Pre-processing: cleaning up the data, removing formatting, detecting the input language, replacing “ä” by “ae” for German, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of *apple* is the right one in a given context?), resolving anaphora and referring expressions like *she*, *the car*, etc.; representing the meaning of the sentence in a machine-readable way

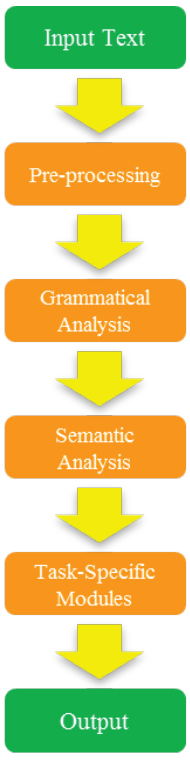


Figure 2: A Typical Text Processing Application Architecture

Task-specific modules then perform many different operations such as automatic summarization of an input text, database look-ups and many others. Below, we will illustrate core application areas and highlight their core modules. Again, the architectures of the applications are highly simplified and idealised, to illustrate the complexity of language technology (LT) applications in a generally understandable way.

After introducing the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of past and ongoing research programs. At the end of this section, we will present an expert estimation on the

situation regarding core LT tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for German.

The most important tools and resources involved are underlined in the text and can also be found in the table at the end of the chapter. The sections discussing the core application areas also contain an overview of the industries active in the respective field in Germany, Austria and Switzerland.

Core application areas

Language Checking

Anyone using a word processing tool such as Microsoft Word has come across a spell checking component that indicates spelling mistakes and proposes corrections. 40 years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognizing syntax–related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in 'She **write* a letter.' However, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e.g., for deciding if a word needs to be written in upper case in German, as in:

Sie übersetzte den Text ins Englische.

[She translated the text into English.]

Er las das englische Buch.

[He read the English book.]

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *englische Buch* is a much more probable word sequence than *Englisch Buch*. A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to German with its flexible word order, compound building, and richer inflection.

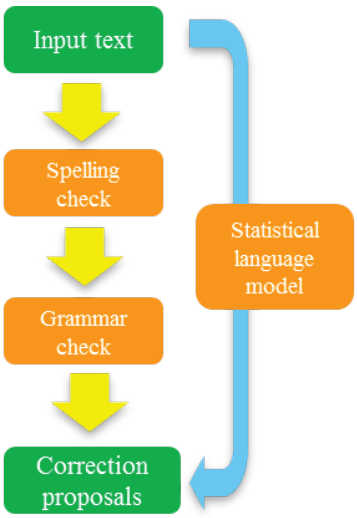


Figure 3: Language Checking (left: rule-based; right: statistical)

The use of Language Checking is not limited to word processing tools, but it is also applied in authoring support systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and dam-

age claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions.

Only few German companies and Language Service Providers offer products in this area. Siemens investigated on German and developed the *Siemens-Dokumentationsdeutsch*[21], a controlled language for German. The German research institute IAI developed a checking module CLAT for German grammar and style. The German company Acrolinx offers a software with a highly adaptable language checker and a terminology database and checking facility. Their style guidelines for technical documentation advise, e.g., against using complex noun compounds like *Achsmesshebebühne* (hydraulic platform for measuring axles) and metaphorical language, e.g., *blitzschnell* (fast as lightning) or *Faustregel* (rule of the thumb). The guidelines also discourage the use of the impersonal pronoun *man* (one, e.g. in *Danach stellt man die Maschine aus.*, lit.: afterwards, one switches off the engine) and of long and nested sentences.

Besides spell checkers and authoring support, Language Checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to Web Search engines, e.g. Google's 'Did you mean…' suggestions.

Web Search

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped language technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide[22]. Since 2004, the verb *googeln* even has an entry in the German *Duden* dictionary. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix[23], which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

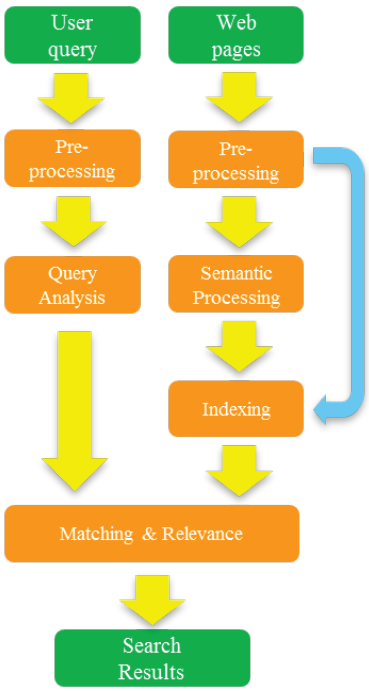


Figure 4: Web Search Architecture

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet (or the equivalent German GermaNet), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *Atomkraft*, *Kernenergie* and *Nuklearenergie* (atomic energy, atomic power, and nuclear energy) or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query 'Give me

a list of all companies that were taken over by other companies in the last five years'. For a satisfactory answer, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This is commonly referred to as information retrieval and involves the search for and ranking of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognizers.

Even more demanding is the attempt to match a query to documents written in a different language. For cross-lingual information retrieval, we have to automatically translate the query to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling multimedia information retrieval, i.e., information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

In Germany, SMEs like Neofonie successfully develop and apply search technologies. The founders of Neofonie provided the first German Web Search engine called *Fireball* in 1997, which was later bought and developed further to a content portal by Lycos Europe. As of today, only few German companies, like Neofonie or Attensity Group, (formerly Empolis), are still providing self-developed search engines. Instead, open source based technologies like Lucene and SOLr are often used by search-focused companies to provide the basic search infrastructure. Other search-based companies rely on international search technologies like, e.g., FAST or Exalead.

Focus on development for these companies lies on providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small text corpora. Processing time easily exceeds that of a common statistical search engine as, e.g., provided by Google by a magnitude of thousands. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.

A meta search engine run by the University of Hannover is MetaGer. Other companies like the Munich-based Intrafind have specialised on search in intranets and standard applications like SAP, where adjustments to the specific customer data is needed. In Switzerland, Eurospider offers information search for internet portals.

In Austria, there are some Web Search engines that are directed to Austrian sites only, such as AT:SEARCH, AUSTRIA-SEEK or AUSTROLINKS. Their coverage and outreach, however, is fairly limited. Apart from these, there are Austrian companies that develop special purpose search engines such as 123people, a real time people search engine with regional searches from Austrian, Ger-

man, Canadian, US, UK, etc. sites as well as world-wide search, or tripwolf, a travel online platform with sites in German and other languages.

Speech Interaction

Speech Interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e.g., a graphical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, and telecommunications. Other usages of Speech Interaction technology are interfaces to particular devices, e.g. in-car navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smartphones.

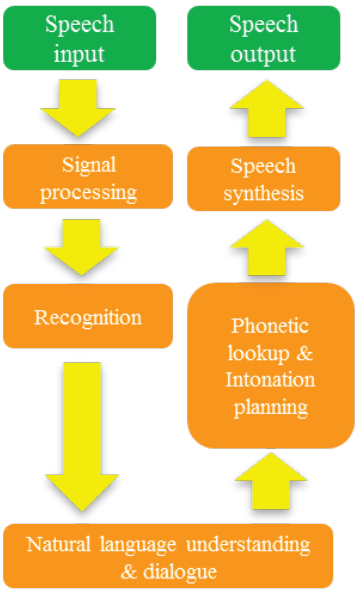


Figure 5: Simple Speech-based Dialogue Architecture

At its core, Speech Interaction comprises the following four different technologies:

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpreting the latter according to the purpose of the respective system.
- Dialogue management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- Speech synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be output to the user.

One of the major challenges is to have an ASR system recognise the words uttered by a user as precisely as possible. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the manual creation of language models that cover a large range of natural language user utterances. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a 'How may I help you' greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible directed dialogue approach.

For the output part of a VUI, companies tend to use pre-recorded utterances of professional – ideally corporate – speakers a lot. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the more dynamic content an utterance needs to consider, the more the user experience may suffer from a poor prosody resulting from concatenating single audio files. In contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for Speech Interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with Nuance and Loquendo being the most prominent ones in Europe.

On the German TTS market, there are additional smaller companies like SVOX, headquartered in Switzerland, voiceINTERconnect and Ivona. An Austrian German TTS voice was commercialized by the UK company CereProc in 2010. For many years, there existed with Philips Speech Recognition Systems a strong ASR research and development unit in Austria, which was acquired by Nuance in 2008. Today, Simon listens is an Austrian non-profit organization developing open source ASR software with a focus on applications for user groups with specific demands such as physically handicapped people and the elderly.

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Today's key players in Germany are Crealog, Excelsis and SemanticEdge. Rather than exclusively relying on a product business based on software licenses, these companies have positioned themselves mostly as full-service providers that offer the creation of VUIs as a system integration service. Finally, within the domain of Speech Interaction, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

As for the actual employment of VUIs, demand in Germany has strongly increased within the last 5 years. This tendency has been driven by end customers' increasing demand for customer self-service and the considerable cost optimisation aspect of automated telephone services, as well as by a significantly increased acceptance of spoken language as a modality for man-machine interaction. These factors were catalysed by the creation of the voice-community.de network, bringing together industry players, research institutes and enterprise customers. Among others, the voice-community initiated a joint elaboration of quality principles for VUIs and organised the annual Voice Days event. These events included a competition for the Voice Awards in different categories. As academic partners, the DFKI and the Fraunhofer IAO institutes were strongly participating in this process of spreading the knowledge about the advantages of Speech Interaction among German enterprises.

Looking beyond today's state of technology, there will be significant changes due to the spread of smartphones as a new platform for managing customer relationships – in addition to the telephone, internet, and email channels. This tendency will also affect the employment of technology for Speech Interaction. On the one hand, demand for telephony-based VUIs will decrease, on the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smartphones will gain significant importance. This tendency is supported by the observable improvement of speaker-independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smartphone users. Given this 'outsourcing' of the recognition task to the infrastructure of applications, the application-

specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

Machine Translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, Machine Translation (MT) still fails to fulfil the high expectations it gave rise to in its early years.

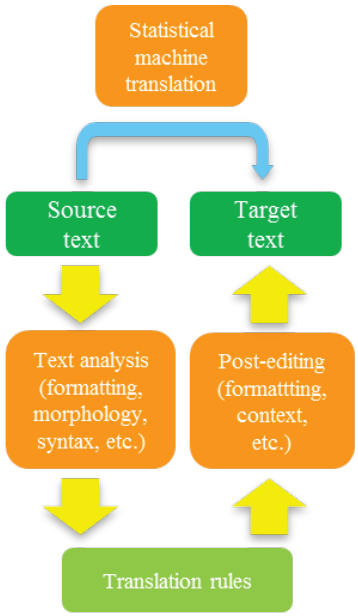


Figure 6: Machine translation (top: statistical; bottom: rule-based)

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g., weather reports. However, for a good translation of less standardized texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty here lies in the fact that human language is ambiguous, which yields challenges on multiple levels, e.g., word sense disambiguation on the lexical level ('Jaguar' can mean a car or an animal) or the attachment of prepositional phrases on the syntactic level as in:

Der Polizist beobachtete den Mann mit dem Fernglas.

[The policeman observed the man with the telescope.]

Der Polizist beobachtete den Mann mit dem Revolver.

[The policeman observed the man with the revolver.]

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like the example above. But often rule-based (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by a skilled linguist.

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in statistical models for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 11 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven MT are complementary, researchers nowadays unanimously target hybrid approaches combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex,

as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For German, MT is particularly challenging. The possibility of creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; free word order and split verb constructions pose problems for analysis, and extensive inflection is a challenge for generating words with proper gender and case markings.

Leading MT systems like LOGOS, METAL (Siemens) and LMT (IBM Heidelberg) were developed in Germany and brought to market maturity. When the big companies ended their engagement, the systems were further developed by offspring and spin-off companies: LOGOS was put open source; METAL was further developed by GMS and later by Lucy Software, and offered in the retail market as Langenscheidt' T1; and the IBM system forms the basis of the product offers of Linguatec ('Personal Translator ') and Lingenio ('translate'). In Switzerland, MT is offered by CLS Communication. All these systems are rule-based. While there is significant research in this technology in national and international contexts, data-driven and hybrid systems have been less successful in business than in research so far.

Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Special systems for interactive translation support were developed, e.g., at Siemens. Language portals, such as the one of Volkswagen, provide access to dictionaries and company-specific terminology, translation memory and MT support.

The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, most of the current systems are English-centred and support only few languages from and into German, which leads to frictions in the total translation workflow, and e.g. forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns allow for comparing the quality of MT systems, the various approaches and the status of MT systems for the different languages. Table 1, presented within the EC Euromatrix+ project, shows the pairwise performances obtained for 22 official EU languages (Irish Gaelic is missing) in terms of BLEU score[24].

The best results (shown in green and blue) were achieved by languages that benefit from considerable research efforts, within coordinated programs, and from the existence of many parallel corpora (e.g. English, French, Dutch, Spanish, German), the worst (in red) by languages that did not benefit from similar efforts, or that are very different from other languages (e.g. Hungarian, Maltese, Finnish).

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Table 1: Pairwise performances obtained for 22 official EU languages in Machine Translation (source: Euromatrix+)

Language Technology

Building language technology applications involves a range of sub-tasks that do not always surface at the level of interaction with the user, but provide significant service functionalities 'under the hood' of the system. Therefore, they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer: 'At what age did Neil Armstrong step on the moon?' - '38'. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled, how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?), and how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context.

This is in turn related to the information extraction (IE) task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. IE aims at identifying specific pieces of information in specific classes of documents; this could be e.g. the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas, which sometimes play the role of standalone application and sometimes that of supportive, 'under the hood' component are text summarization and text generation. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying 'important' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarization equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarizers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but embedded into a larger software environment, such as into the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

For German, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarization have since the 1990s been the subject of numerous open competitions, primarily those organized by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but German was never prominent. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most available software is for English. There is a German version of one semantics-based multilingual realizer, as well as a template-based German realizer; however, both originated in the 1990s and have not been ported to contemporary software environments.

Language Technology in Education

Language technology is a highly interdisciplinary field, involving the expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists, among others. As such, it has not yet acquired a fixed place in the German faculty system. Some universities have established a separate institute of Computational Linguistics (CL), e.g. Heidelberg, Saarbrücken, and Tübingen, sometimes under a slightly different name (e.g. in Stuttgart). However, programmes are also offered by other departments, such as the faculty of computer science (e.g. in Leipzig and Hamburg) or the faculty of linguistics (e.g. in Bochum and Jena). Some universities offer master courses only (e.g. Gießen), bachelor courses only (e.g. Erlangen-Nürnberg, Göttingen, Munich, Potsdam, Trier) or modules in language technology to students of other courses of study (e.g. Hildesheim). Many of these programs and courses have only recently been introduced. Currently, at least 17 German universities offer programs in the field of language technology. In Switzerland, CL programs are offered by the

Universities of Zurich and Geneva. There is no regular CL study programme in Austria, but CL- and LT-related courses are taught as part of other studies, mainly in Vienna, but also in Klagenfurt.

The German Federal Statistical Office has been recording statistics on CL as a course of study at German universities since winter semester 1992/1993. In the following years, studying CL became increasingly popular. Since the turn of the century, the number of students has been stable with annually around 250 – 350 new students enrolling for CL as their main course of study.[25] With this comparably low number of students, the steadily rising demand of qualified personnel specialized in the field of language technology cannot be met by graduates from German universities alone. In many cases, companies and research institutes, such as the German Research Center for Artificial Intelligence (DFKI) and the Austrian Research Institute for Artificial Intelligence (ÖFAI) need to draw upon experts from abroad.

Language Technology Programs

The existence of a comparably vivid LT industry in Germany can be traced back to major LT programs carried out in the last decades. One of the first such programs was EUROTRA, an ambitious Machine Translation (MT) project established and funded by the European Commission from the late 1970s until 1994. Even though the EUROTRA project did not fulfil the expectations of creating a state-of-the-art MT system, the project had a long-term impact on the language industries in Europe. Due to a shift in the translation paradigm, more data-driven solutions were explored in the VERBMOBIL project. This large national project on translation of spontaneous speech between German, Japanese and English was funded by the Federal Ministry of Education and Research (BMBF) between 1993 and 2000. The VERBMOBIL prototype itself was not able to establish on the market, but has lead to many other innovations and, today, forms the basis of the Google Translate system.

The IBM project LILOG (1985-1991) was an implementation of an information base in the German language. It involved approximately 200 of the scientists working in Germany in the fields of Computational Linguistics, natural language understanding systems, and Artificial Intelligence and showed that a cooperative project between universities and industry can produce useful results both in pure research and in implemented methods and tools.

National projects on the mark-up and annotation of language resources were funded in the 1990s and early 2000 and led to the development of the Stuttgart-Tübingen tagset (STTS), which has had a lasting impact on current work on annotation of language corpora. Two further projects were NEGRA and TIGER, both partially funded by the German Research Foundation (DFG). The annotation schemes proposed by those projects have become de facto standards in the field and some adaptation and abstraction work over those schemes has brought about an international standardisation of syntactic annotation.

COLLATE (BMBF, 2000-2006) was one of the first projects to address the issues of a language infrastructure, including the creation of an information portal about the field (LT World). German and Austrian institutions are also involved in the ongoing CLARIN project. Other ongoing projects include those comprised by EUROPEANA, and THESEUS , a project co-funded by the Federal Ministry of Economics and Technology (BMWI) with the goal to develop

the basic technologies and standards necessary to make knowledge on the Internet more widely available in the future.

Together with many lower-scale projects funded and carried out, the mentioned projects have led to the development of competence in the field of language technology as well as a basic technological infrastructure of language tools and resources for German. Still, public funding for LT projects in Germany and in Europe is relatively low compared to the expenses spent on issues like language translation and multilingual information access by the USA[26].

In Austria, the Medical University of Vienna developed a language dialog system in German within the VIE-LANG project. The Faculty of Computer Sciences of the University of Vienna is conducting the JETCAT project on Japanese-English translation. Since 2001, there has been an on-going project to produce the Austrian Academy Corpus. There are no specifically dedicated funding programmes for LT in Austria. Funding for LT-related topics typically comes from research programmes with open topics and especially programmes with a special focus on the transfer from academic research to industry, in particular SMEs. Several of these programmes are administered by the Austrian Research Promotion Agency (FFG). The Vienna Science and Technology Fund (WWTF) is a comparably strong supporter of localized language technology, especially topics with a strong relation to Vienna such as the Viennese dialect and sociolect synthesis and Machine Translation from Austrian German to Viennese and other dialectal variants.

In Switzerland, interest in language technology began in the 1980s, with a strong involvement in the EUROTRA project. Currently, the Universities of Zurich and Geneva conduct several projects in the field of MT, including MT between Standard German and Swiss German[27]. Corpus building projects include the collection of speech corpora by the National Centre of Competence in Research on Interactive Multimodal Information Management[28] and a project that collects SMSes in Swiss German.[29] Suisse research institutes in the field include ISSCO and IDIAP. Overall, the Suisse LT scene is small, mainly due to the limited funding opportunities. EU funding is not always accessible and unattractive for Suisse SMEs. On the other hand, the Commission for Technology and Innovation (KTI) offers efficient and unbureaucratic support to short- and medium-term projects. It also supports the development of start-up companies. However, due to a lack of experts, start-ups in the field of language technology are rare.

Availability of Tools and Resources for German

The following table provides an overview of the current situation of language technology support for German. The rating of existing tools and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

1. **Quantity**: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.
   - 0: no tools/resources whatsoever
   - 6: many tools/resources, large variety
2. **Availability**: Are tools/resources accessible, i.e.,are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?

- 0: practically all tools/resources are only available for a high price
- 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing

3 **Quality**: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?

- 0: toy resource/tool
- 6: high-quality tool, human-quality annotations in a resource

4 **Coverage**: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?

- 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
- 6: very broad coverage resource, very robust tool, widely applicable, many languages supported

5 **Maturity**: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.

- 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
- 6: immediately integratable/applicable component

6 **Sustainability**: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

- 0: completely proprietary, ad hoc data formats and APIs
- 6: full standard-compliance, fully documented

7 **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?

- 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
- 6: very high level of adaptability; adaptation also very easy and efficiently possible

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	2	4	4	4	3	3
Parsing (shallow or deep syntactic analysis)	4	3	3	3	4	2	2
Sentence Semantics (WSD, argument structure, semantic roles)	2	2	2	1	2	1	2
Text Semantics(coreferenceresolution, context, pragmatics, inference)	2	1	3	2	2	2	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	3	2	3	2	2	2	1
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	4	4	3	3	5	5	5
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	2	3	2	4	3	2
Language Generation (sentence generation, report generation, text generation)	2	1	2	2	2	1	2
Summarization, Question Answering,advanced Information Access Technologies	2	2	2	2	3	1	1
Machine Translation	5	3	2	3	4	1	2
Speech Recognition	5	1	4	4	4	3	3
Speech Synthesis	5	3	4	5	4	3	3
Dialogue Management (dialogue capabilities and user modelling)	3	2	4	3	4	5	5
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	1	4	3	5	5	3
Syntax-Corpora(treebanks, dependency banks)	3	3	3	2	3	3	2
Semantics-Corpora	2	2	2	1	2	2	2
Discourse-Corpora	1	2	2	2	1	1	1
Parallel Corpora, Translation Memories	2	1	2	2	2	2	1
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	3	1	3	2	3	3	2
Multimedia and multimodal data	2	1	3	1	1	2	1

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
(text data combined with audio/video)							
Language Models	2	1	3	3	3	1	1
Lexicons, Terminologies	4	2	4	3	4	4	2
Grammars	3	2	3	3	3	2	1
Thesauri, WordNets	2	3	3	1	4	4	3
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	3	3	2	3	3	1	1

This Table on the status of Technologies and Resources (Data, Tools, Evaluation and Meta-resources) for the German language is close to the one produced for French. The situation is actually very similar, as can also be seen in the Euromatrix+[30] Bilingual Tables. In the META-Matrixes, produced by META-NET from the data obtained in the LRE Map[31], it appears that, among the 23 EU official languages, French and German get about the same number of resources overall (respectively 143 and 132), far from what exists for English (559), and followed by Spanish (111) and Italian (90). In the Euromatrix+, produced from the Hutchins Compendium of Translation Software[32], French and German are also close (130 and 140 respectively), far from English (257), and followed by Spanish (128) and Italian (116).

For German, key results regarding technologies and resources include the following:

- Speech processing currently seems to be more mature than processing of written text. Advanced information access technologies are in their infancies.
- The more linguistic and semantic knowledge a tool takes into account, the more gaps exist (see, e.g., information retrieval vs. text semantics); more efforts for supporting deep linguistic processing are needed.
- Research was successful in designing particular high quality software, but many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.
- For German, a large reference text corpus (with a balanced mixture of various genres) exists, but it is not easily/cheaply accessible.
- Annotated corpora with syntactic, semantic, or even discourse structures are missing; again, the situation is worse the more deep linguistic and semantic information is needed.

- Parallel corpora for machine translation are missing. As can be seen in Table 1, translation of German into English works best, as most data exists.
- Multimedia data is a huge gap.

Conclusions

In this Whitepaper Series, the first effort has been made to assess the overall situation of many European languages with respect to Language Technology support in a way that allows for high level comparison and identification of gaps and needs.

The situation of German concerning Language Technology support gives rise to cautious optimism. Supported by larger research programs in the past, there exists a Language Technology industry and research scene in Germany, Austria, and Switzerland. A lot of large size resources and state-of-the-art technologies have been produced and distributed standard German. However, the size of the resources and the number of tools are still very limited compared to what exists for the English language, and still insufficient to address all the technologies related to German that are needed for offering the support a true multilingual knowledge society needs.

It is evident that technologies that have been developed and optimised for the English language in many cases do not easily carry over to German. For example, identical systems for parsing (syntactic and grammatical analysis of the sentence structure) typically show a much lower performance on German text as compared to English in international competitions. The reasons lie in the special characteristics of German such as free word order or long and nested sentences that require more sophisticated processing.

Unfortunately, the language technology industry working on German is limited, and most of the large European companies have ceased or decreased their activity in that area leaving the field to several SMEs, which can hardly attack an international market while the language barrier appears as one of the main factor for limiting cross-border e-Commerce in the EU33.

From this, it is clear that more efforts need to be directed into the creation of resources for German and into research, innovation, and development. The need for large amounts data and the high complexity of Language Technology systems make it also mandatory to develop new infrastructures for sharing and cooperation.

The R&D funding lacks continuity, with short term coordinated programs interrupted by periods of low and sparse funding, and missing coordination with other programs existing in other EU countries or at the European Commission.

A large, coordinated effort on Language Technologies would help saving the German language just like the other languages, and multilingualism in general in Europe and worldwide34.

About META-NET

META-NET is a Network of Excellence funded by the European Commission. The network currently consists of 47 members from 31 European countries. META-NET fosters the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

The Multilingual Europe Technology Alliance (META)

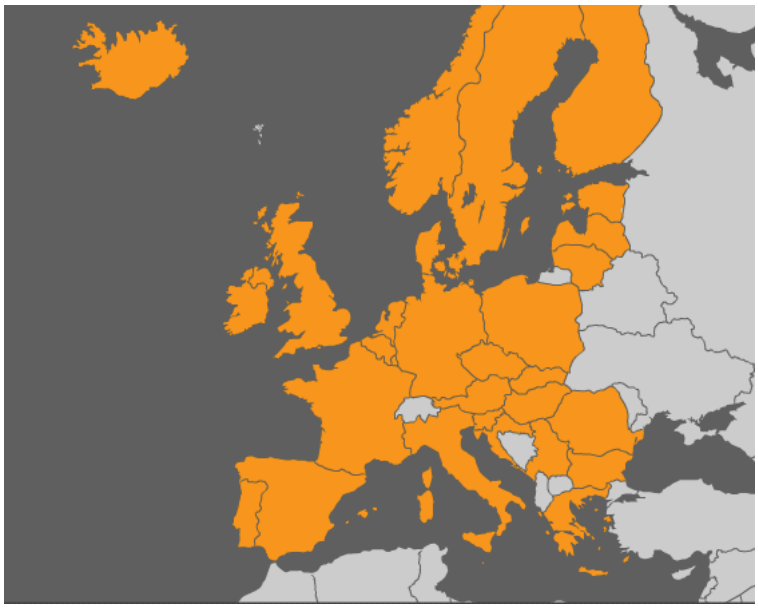


Figure 1: Countries Represented in META-NET

META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- provides equal access to information and knowledge in any language;
- offers advanced and affordable networked information technology to European citizens.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

Lines of Action

META-NET launched on 1 February 2010 with the goal of advancing research in language technology (LT). The network supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its

goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.

Figure 2: Three Lines of Action in META-NET

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In the first year of META-NET, presentations at the FLaReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT—from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collect-

ing data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

Member Organisations

The following table lists the organisations and their representatives that participate in META-NET.

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	University of Zagreb	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Republic	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bolette Sandford Pedersen and Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi and Krister Linden
France	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Germany	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	Saarland University	Manfred Pinkal
Greece	Institute for Language and Speech Processing, "Athena" R.C.	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi

Country	Organisation	Participant(s)
	Budapest University of Technology and Economics	Géza Németh and Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Utrecht University	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski and Maciej Ogrodniczuk
	University of Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pęzik
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	Alexandru Ioan Cuza University	Dan Cristea
Serbia	University of Belgrade	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institute Mihailo Pupin	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	Pompeu Fabra University	Núria Bel

Country	Organisation	Participant(s)
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou
	University of Edinburgh	Steve Renals

References

[1] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[2] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[3] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[4] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[5] http://cdt.europa.eu/EN/whoweare/Pages/OurEULanguages.aspx

[6] http://www.goethe.de/mmo/priv/1459127-STANDARD.pdf

[7] http://www.efnil.org/documents/language-legislation-version-2007/germany/germany

[8] http://www.canoo.net/services/GermanSpelling/Reform/fremdwoerter/eindeutschung-lebend.jsp?MenuId=GermanSpellingReform111

[9] Lothar Lemnitzer: *Von Aldianer bis Zauselquote*, Gunter Narr Verlag Tübingen, 2007.

[10] http://www.spiegel.de/thema/zwiebelfisch/

[11] Wolf Schneider: *Speak German! – Warum Deutsch manchmal besser ist*. Rowohlt, 2008.

[12] "PISA 2009 Ergebnisse: Zusammenfassung", at http://www.pisa.oecd.org/dataoecd/34/19/46619755.pdf

[13] http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_01_01-Bildungsbericht-erste-Befunde.pdf

[14] http://userweb.port.ac.uk/~joyce1/abinitio/whygerm7.html

[15] http://userweb.port.ac.uk/~joyce1/abinitio/whygerm7.html

[16] http://www.ard-zdf-onlinestudie.de/

[17] http://www.statistik.at/web_en/statistics/information_society/ict_usage_in_households/041019.html.

[18] http://www.denic.de/hintergrund/geschichte-der-denic-eg.html

[19] http://www.ebrandservices.com/welcome-to-e-brand-services,130.html

[20] http://dict.leo.org/

[21] Anne Lehrndorferand Stefanie Schachtl: "TR09: Controlled Siemens Documentary German andTopTrans", *TC Forum*, 1998.

[22] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[23] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

[24] The higher the score, the better the translation, a human translator would get around 80. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA.

[25] Based on official university statistics of the Federal Statistical Office in Wiesbaden (29.03.2011)

[26] Gianni Lazzari: „Sprachtechnologien für Europa“, 2006: http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf

[27] http://www.latl.unige.ch/personal/yvesscherrer/#talks_papers.fr

[28] Claudia Soria, Joseph Mariani(2011): “Report on Existing Projects and Initiatives”

[29] http://www.sms4science.ch/

[30] http://www.euromatrixplus.net/matrix/

[31] http://www.resourcebook.eu/LreMap/faces/views/resourceMap.xhtml

[32] www.hutchinsweb.me.uk/Compendium-14.pdf

[33] http://ec.europa.eu/consumers/strategy/docs/com_staff_wp2009_en.pdf

[34] V. Reding, J.Figel’, Preface, in *Human Language Technologies for Europe*, TC-Star project, http://www.tcstar.org/pubblicazioni/D17_HLT_ENG.pdf